# Prediction of Thyroid Disease Using Machine Learning

S. Dollar Venkata Ramana deekshith[1], M. Sai Pavan[2], S. Joseph Reddy[3], Sk Sadik[4], S. Ramadoss[5]
[1,2,3,4] *Student/Dept of CSE, School of Computing, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu.*
[5]*AsstProfessor/Dept of CSE, School of computing, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu*

***Abstract*—Thyroid medical diagnosis and prediction development, which medical science is a complicated axiom. Thyroid gland is one of our body's main organs. Thyroid hormone secretions are responsible for regulating metabolism. Hyperthyroidism and hypothyroidism are the two prominent thyroid disorders that produce thyroid hormones for control of body metabolism. The machine learning is critical in the disease prediction process and in the study and classification models used for thyroid disease on the basis of data obtained from hospital datasets. A decent knowledge base must be ensured, built and used as a hybrid model to solve dynamic learning tasks like medical diagnosis and prediction tasks. Basic techniques of machine learning are used for the identification and inhibition of thyroid. The SVM is used to predict the approximate probability of a thyroid patient. If the patient has risk of getting thyroid our system has to give suggestions like recommending home remedies, precautions, medication secretions are responsible for regulating metabolism. Hyperthyroidism and hypothyroidism are the two prominent thyroid disorders that produce thyroid hormones for control of body metabolism. The machine learning is critical in the disease prediction process and in the study and classification models used for thyroid disease on the basis of data obtained from hospital datasets.decent knowledge base must be ensured, built and used as a hybrid model to solve dynamic learning tasks like medical diagnosis and prediction tasks. Basic techniques of machine learning are used for the identification and inhibition of thyroid. The SVM is used to predict the approximate probability of a thyroid patient. If the patient has risk of getting thyroid our system has to give suggestions like recommending home remedies, precautions, medication etc.Thyroid disorders, such as hypothyroidism, hyperthyroidism, and thyroid cancer, pose significant health challenges worldwide**

***Index Terms*—Machine Learning Algorithm, Thyroid disease, Support Vector Machine (SVM), K-NN, Decision Trees Prediction**

## I. INTRODUCTION

Advanced machine biology is used in the area of healthcare. It required data to be collected for medical disease prediction. For early-stage disease detection, various intelligent prediction algorithms are used. The Medical Information System is good with data sets, but intelligent systems are not available for the fast diagnosis of diseases. Eventually, machine learning algorithms play a key position in solving complex and non-linear problems during the creation of prediction model. The characteristics that can be selected from the various data sets that can be used as description in a healthy patient as specifically as possible are needed in any disease prediction models. Otherwise, misclassification can result in a good patient receiving inappropriate care. The reality of forecasting any condition associated with thyroid illness is also of the greatest cardinal number. Thyroid gland is endocrine in stomach. It is erected in lowered portion of human neck, under apple of Adam, and assists in secretion of thyroid hormones and which ultimately affects metabolism rate and protein synthesis. To control body metabolism, these hormones count on how quickly heart beats and how quickly calories burn. The composition of thyroid hormones helps to control the body's metabolism. These glands consist of two mature levothyroxine (abbreviated T4) and triiodothyronine thyroid hormones (abbreviated T3). These thyroid hormones are essential for manufacturing and general construction and regulation in order to regulate body temperature. T4 and T3 are exclusively two activated thyroid hormones that usually compose of thyroid glands. These hormones are vital to the control of proteins; distribution at body temperature and energy- bearing

and propagation in every part of the body. With T3 and T4 hormones, iodine is primary building block of thyroid glands and is prostrate in only some unique problems, which are exceedingly prevalent. Insufficient elements of these hormones to hypothyroidism and an inappropriate portion to hyper thyroidism. Hyperthyroidism and underactive thyroidism have multiple origins. There are a number of drugs.

Thyroid surgery is weak to ionizing radiation, continuous thyroid softness, iodine deficiency, and loss of enzyme to produce thyroid hormones.

Thyroid disorders, including hypothyroidism and hyperthyroidism, are common endocrine diseases that can significantly impact metabolism and overall health. Early detection and accurate diagnosis are crucial for effective treatment and management. Traditional diagnostic methods, such as blood tests for thyroid hormones (TSH, T3, T4) and medical imaging, can be time-consuming and require expert interpretation. With the advancements in artificial intelligence (AI) and machine learning (ML), automated diagnostic systems have gained attention for their ability to enhance accuracy and efficiency. Among various ML algorithms, Random Forest is a highly effective ensemble learning technique used for classification and prediction tasks. It constructs multiple decision trees and combines their outputs to make reliable and robust predictions. In thyroid detection, Random Forest can analyze patient data, including hormone levels and clinical symptoms, to classify whether a patient has a thyroid disorder. Its ability to handle missing data, identify important features, and provide high accuracy makes it a suitable choice for medical diagnosis. This study explores the implementation of Random Forest for thyroid disease detection, highlighting its advantages, working mechanism, and potential impact on healthcare.

This study explores the application of machine learning in thyroid detection, focusing on data preprocessing, feature selection, model development, and performance evaluation. The goal is to improve diagnostic precision, reduce human error, and facilitate early intervention, ultimately enhancing patient outcomes. The implementation of such an intelligent system can assist healthcare professionals in faster, more accurate, and automated diagnosis, reducing the dependency on extensive lab testing and improving patient outcomes. The study explores data preprocessing, model training, performance evaluation, and real-world applicability of the Random Forest-based thyroid detection system.

## II. LITERATURE SURVEY

2.1A modern hybrid approach for thyroid diagnosis based on the artificial immune recognition system (AIRS) with blurry, weighted pre-treatment: Proper understanding of the functional data of thyroid gland is important concern diagnosis of thyroid disease. Major function of thyroid gland is to help control the metabolism of body. This is provided by the thyroid hormone released by thyroid gland. Containing very little thyroid hormone (hypothyroidism) or producing more thyroid hormone (hyperthyroidism) determines kind of thyroid disorder. Artificial Immune Systems are recent yet powerful branch of artificial intelligence. Artificial Immune Recognition System (AIRS) as suggested has so far been one of the systems proposed in this area. Watkins has demonstrated an important and fascinating success on the topics that have been discussed. The purpose of this research is to diagnose thyroid using modern hybrid machine learning process, with classification scheme. Via the hybridization of AIRS with formed Fuzzy, the methods used to solve diagnostic problem by grouping are weighted. The strength of the samples is checked using a cross- validation process. We used the data collection for thyroid disease taken from the UCI respiratory learning machine. We have also reached 85% classification accuracy, which is highest achieved to date. Accuracy of the classification was achieved it is a having 10-fold cross-validation. disorders. The problem is sometimes compounded in elderly patients whose signs are often masked or due to other medical conditions. While lab-oratory tests have become more reliable and are effective in diagnosing thyroid disorders (positive predictive rates of certain tests have recently been reported to be over 90%. Findings are also not very satisfactory across all cases. Diagnostic problems emerge from variability in test results across patients and other considerations such breastfeeding, interactive medications. 2.4 A survey on applying machine learning techniques for management of diseases:

2.3 In recent years, increasing research knowledge and large data output have contributed to exponential rises in databases and repositories. Biomedicine is one of rich data domains. Currently, a wide range of biomedical information available, ranging from explanations of clinical disorders to various types of biochemical data and image instrument outputs. Manually extracting and converting biomedical trends from data into machine-understood information is daunting task as the biomedical domain consists of vast, dynamic and nuanced expertise. Data mining can improve biomedical pattern extraction performance. A description of the uses of data mining for is not in an i disease control is provided in this report. The key emphasis analyzes machine learning techniques (MLTs) that commonly used forecast, predict, and treat major common diseases such as cancer, hepatitis, and heart disease. The methods, including the Artificial Neural Network, K- Nearest Neighbor, Decision Tree and Associative Grouping, are explained and analyzed. This survey offers general overview of current state of MLT disease control. The precision obtained for the different applications varied from 70% to 100% based on the disorder, the problem solved, and the data and procedure used.

2.4 Machine learning has significantly improved thyroid disease detection, offering high accuracy and efficiency. However, challenges such as data quality, model interpretability, and real- time deployment need to be addressed for broader clinical adoption. Future advancements in deep learning and hybrid approaches are expected to enhance diagnostic accuracy further. Cloud-based AI models, explainable AI, and better real-time applications in clinical settings. Combining multiple ML models, such as hybrid deep learning models (CNN-RNN), has demonstrated improved accuracy in thyroid diagnosis. Ensemble methods like XGBoost and AdaBoost enhance classification performance by reducing overfitting and improving generalization. Support Vector Machines (SVM): Studies have shown SVM to be effective in classifying thyroid disease, achieving high accuracy in distinguishing normal and abnormal thyroid function based on lab results.

2.5 Decision Trees and Random ForestsThese models are useful for feature selection and classification, providing interpretable results. Artificial Neural Networks (ANN): ANNs have been widely used for thyroid disease classification, leveraging multiple layers for feature extraction and pattern recognition.k-Nearest Neighbors (k-NN): This algorithm has been applied in thyroid disease classification but is often less accurate compared to other methods.

## III. EXISTING SYSTEM

Thyroid disorder significant cause of formation of medical diagnosis and estimation, which is a challenging axiom of medical science. The secretions of thyroid hormones are guilty of metabolism regulation. Hyperthyroidism and hypothyroidism are one of the two prevalent thyroid disorders that release thyroid hormones to control body metabolism. Data cleaning methods have been used to make data primitive enough to do analytics to demonstrate likelihood of patients having thyroid.

DISADVANTAGES

- Energy level
- Weakness
- Breathing

## IV. PROPOSED SYSTEM

In the prediction process, machine learning plays a key role, and paper research and the classifications of models used in thyroid disease are based on information from UCI machine learning repositories. A decent knowledge base that can be centered and used as a hybrid paradigm must be preserved in order to address complex learning issues, such as medical diagnostics and statistical tasks. We also proposed different approaches for machine learning and thyroid diagnosis. Machine Learning Algorithms, Vector Support Machine, were used to calculate an estimated probability of a patient having thyroid disease.

ADVANTAGES

- Avoids long-term risks of anti-thyroid and radioactive iodine.
- Provides histology tissue, for childbearing instantly.

Data Collection

- The dataset consists of patient records with features such as TSH (Thyroid-Stimulating Hormone), T3, T4, age, gender, and other clinical attributes.
- Publicly available datasets like UCI Thyroid Disease Dataset can be used for training.

Data Preprocessing

- Handling missing values using imputation techniques.
- Normalization and standardization of numerical features.
- Encoding categorical data (e.g., gender, symptoms).
- Feature selection to retain the most relevant attributes.

Model Training using Random Forest

- The dataset is split into training and testing sets (e.g., 80%-20%).
- The Random Forest classifier is trained using multiple decision trees to improve generalization and reduce overfitting.
- Hyperparameter tuning (number of trees, max depth, etc.) is performed to optimize performance.

Model Evaluation

- The trained model is evaluated using accuracy, precision, recall, and F1-score.
- Confusion matrices and ROC-AUC curves are used to assess classification performance.

Prediction & Deployment

- The trained model is deployed as a web- based or mobile application for real-time thyroid disease prediction.
- Users input their test results, and the system provides a classification along with confidence scores.
- The proposed system aims to develop an efficient thyroid disease detection and classification model using the Random Forest (RF) algorithm. Random Forest is an ensemble learning method that improves classification accuracy by combining multiple decision trees.
- The system will classify patients as hypothyroid, hyperthyroid, or normal based on medical data, such as T3, T4, and TSH levels. Compared to traditional decision tree
- Compared to traditional decision trees and statistical models.
- Compared to traditional decision trees and statistical models.
- Compared to traditional decision trees.

networks can have good estimations of posterior probabilities and thus provide improved classification efficiency than conventional statistical approaches such as logistic regression. Furthermore, neural network models have seen to be resilient to sampling variations. It is demonstrated for medical diagnosis issues, where data sometimes very unbalanced, neural networks may be a promising classification tool for practical use. Proper diagnosis of thyroid dysfunction based on clinical and experimental testing is often complicated. One explanation is the non- specific existence of certain signs of thyroid disease. This is particularly true in hypothyroidism, where signs such as lethargy, confusion, weight gain, impaired memory are easily associated with psychological and medical disorders. The problem is sometimes compounded in elderly patients whose signs are often masked or due to other medical conditions. While lab-oratory tests have become more reliable and are effective in diagnosing thyroid disorders (positive predictive rates of certain tests have recently been reported to be over 90%. Findings are also not very satisfactory across all cases. Diagnostic problems emerge from variability in test results across patients and other considerations such breastfeeding, interactive medications.

A survey on applying machine learning techniques for management of diseases:

In recent years, increasing research knowledge and large data output have contributed to exponential rises in databases and repositories. Biomedicine is one of rich data domains. Currently, a wide range of biomedical information available, ranging from explanations of clinical disorders to various types of biochemical data and image instrument outputs. Manually extracting and converting biomedical trends from data into machine- understood information is daunting task as the biomedical domain consists of vast, dynamic and nuanced expertise. Data mining can improve biomedical pattern extraction performance. A description of the uses of data mining for disease control is provided in this report. The key emphasis

analyzes machine learning techniques (MLTs) that commonly used forecast, predict, and treat major common diseases such as cancer, hepatitis, and heart disease. The methods, including the Artificial Neural Network, K- Nearest Neighbor, Decision Tree and Associative Grouping, are explained and analyzed. This survey offers general overview of current state of MLT disease control. The precision obtained for the different applications varied from 70% to 100% based on the disorder, the problem solved, and the data and procedure used.

## V. MODULES

1. Data Collection:

In this project we are using SVM machine learning algorithm to predict whether patient report data is normal or at risk of thyroid disease and if thyroid disease predicted from patient report, then application will display proper diet and medication details. In this project we are using UCI machine learning THYROID disease dataset to train SVM algorithm and to generate prediction model. New patient test data will be applied on trained SVM model to predict whether patient is normal or at risk of thyroid disease.

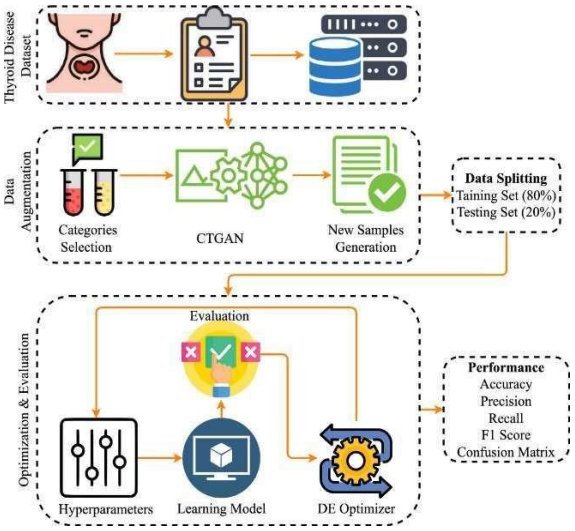
2. Preprocessing:

In Data cleaning the system detect and correct corrupt or inaccurate records from database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying or detecting the dirty or coarse data. In Data processing the system converts data from a given form to a much more usable and desired form i.e. makes it more meaningful and informative. Post processing procedures usually include various pruning routines, rule quality processing, and rule filtering, rule combination, model combination, or even knowledge integration. All these procedures provide a kind of symbolic filter for noisy, imprecise, or non-user-friendly knowledge derived by an inductive algorithm

3. Feature Extraction:

Data Typically, Here the system separate a dataset into a training set and testing set, most of the data use for training, and a smaller portion of data is use for testing. After a system has been processed by using the training set, it makes the prediction against the test set.

SYSTEM ARCHITECTURE:



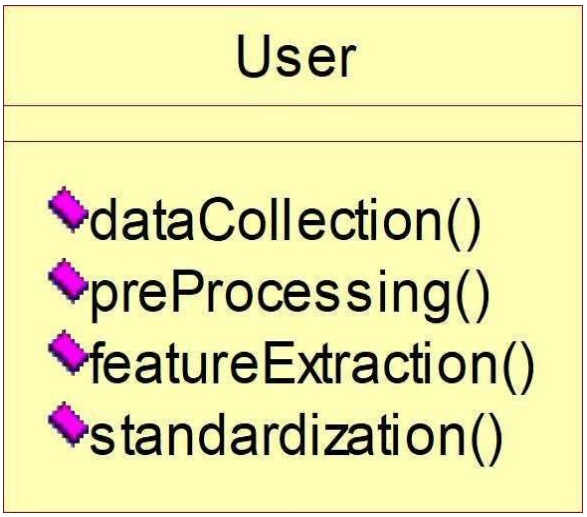

4. UML DIAGRAMS

The three major elements of UML are

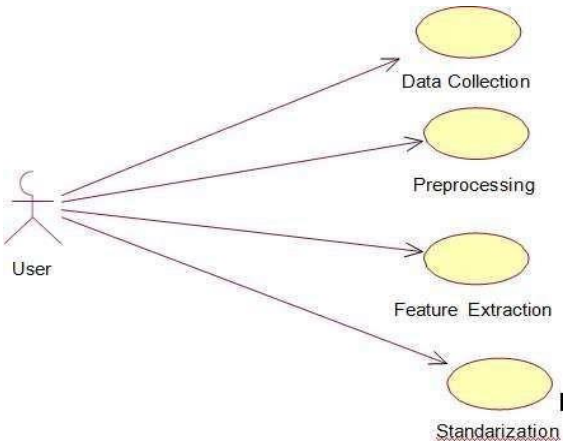
1. The UML's basic building blocks
2. The rules that dictate how those building blocks may be put together
3. Some common mechanism that applies throughout the UML.

1. CLASS DIAGRAM:

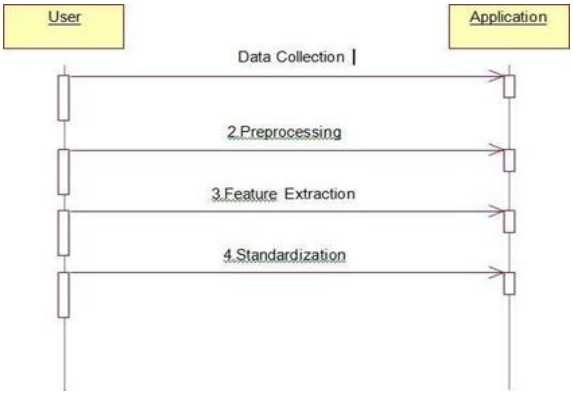
In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.



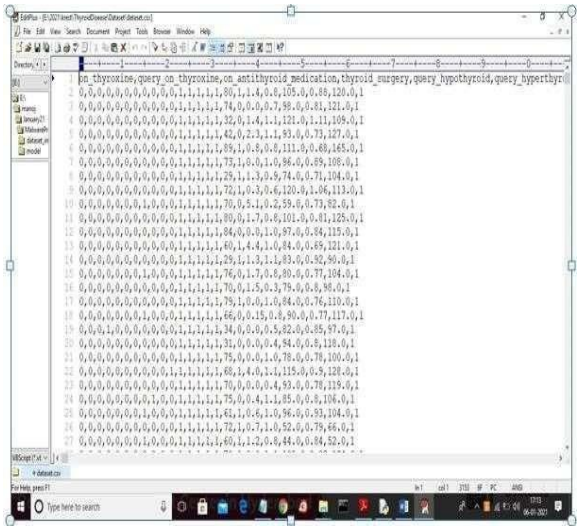
2. USE CASE DIAGRAM:



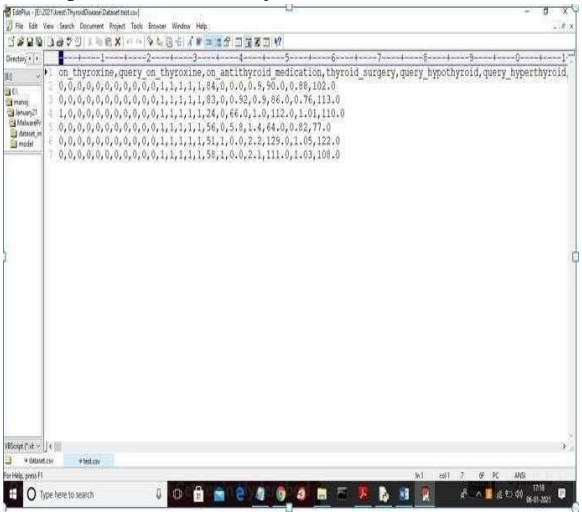

3. SEQUENCE DIAGRAM:




4. OUTPUT SCREENS:



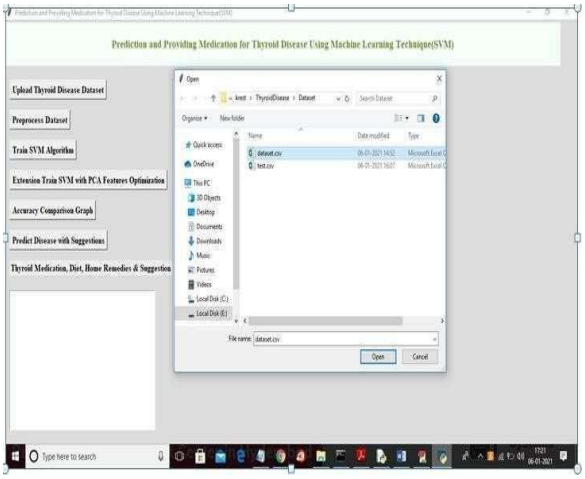
In above dataset first row contains column names and other rows contains values as 0 or 1 and if patient is under thyroid medication or surgery then its column value will be 1 else 0 and in last column contains class label as 0 or 1 where 0 means patient record is normal and 1 means patient record contains thyroid disease. In this dataset more than 3000 rows are there and 24 columns are available. For prediction all 24 columns are not available so we are applying PCA (principal component analysis) feature selection algorithm as extension concept to optimize features or to reduce columns or features which are not important for prediction. PCA will remove unnecessary columns from the dataset and use only important attributes to train SVM algorithm and due to optimize features SVM prediction accuracy can be increase.



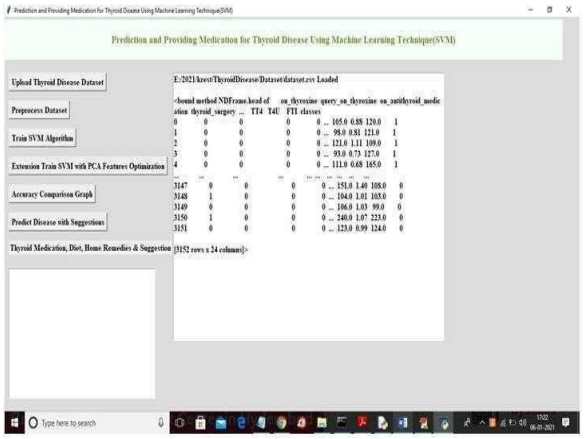
In above test dataset we can see there is no class label with value 0 or 1 and SVM will predict that value

To run project double click on 'run.bat' file to get below screen

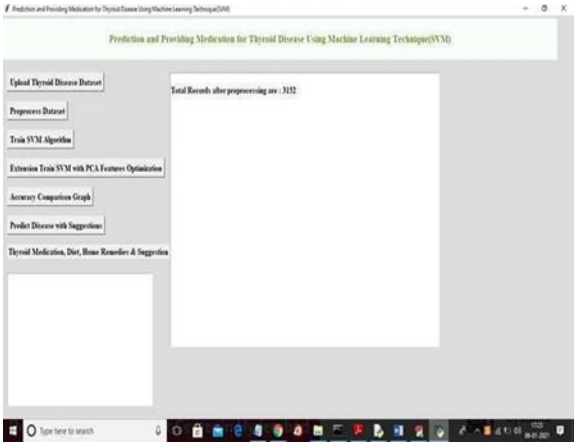

In above screen click on 'Upload Thyroid Disease Dataset' button to upload dataset and to get below screen



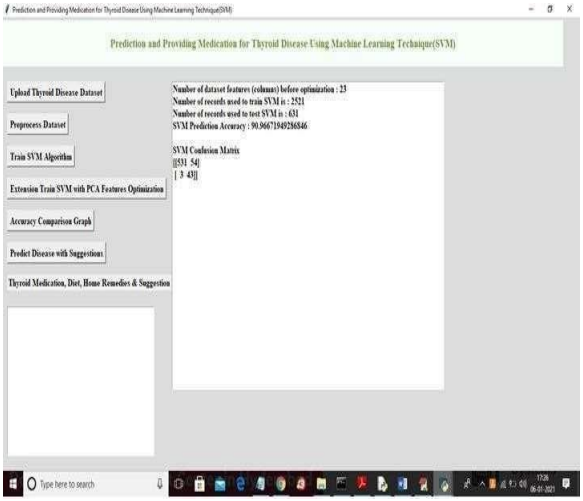

In above screen selecting and uploading 'dataset.csv' file and then click on 'Open' button to load dataset and to get below screen



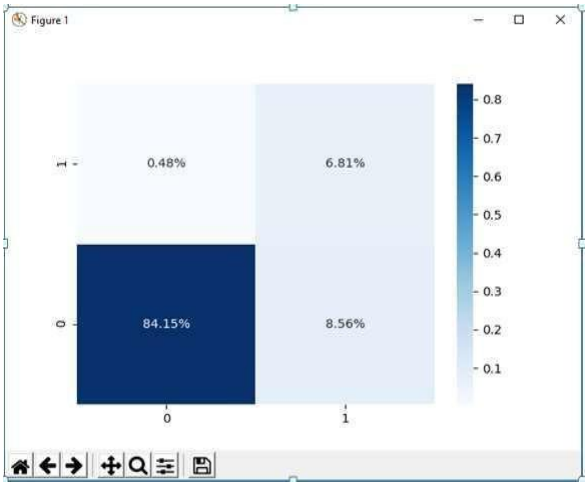

In above screen dataset loaded and displaying few records from dataset and then click on 'Preprocess Dataset' button to remove missing and NAN values from dataset and to separate X and Y values where X contains all dataset values and Y contains class label value.



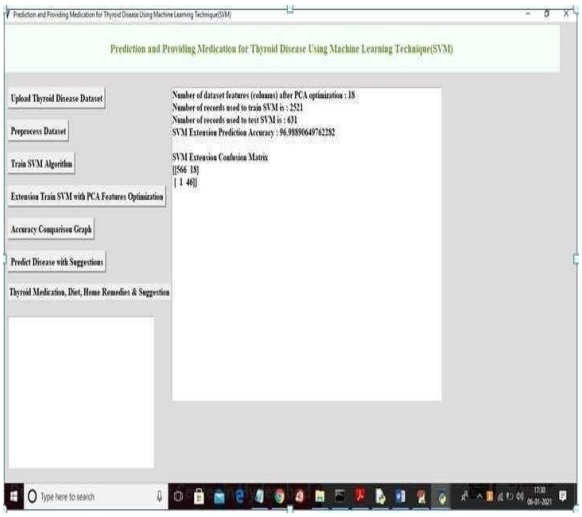

In above screen dataset showing 3152 preprocess records and now dataset is ready and now click on 'Train SVM Algorithm' button to split dataset into train and test and then apply SVM algorithm on train data to generate model and then model will be applied on test data to calculate prediction accuracy



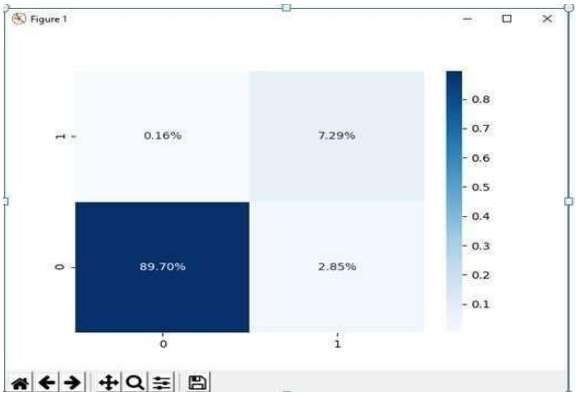

In above screen we can see dataset contains total 23 columns and using 2521 records to train SVM algorithm and using 631 test records to test SVM prediction accuracy and with normal SVM we got prediction accuracy as 90.96% and application showing confusion matrix of true and false prediction values where 531 and 3 are the true prediction and 54 and 43 are the false or incorrect prediction and below is the graph format of confusion matrix



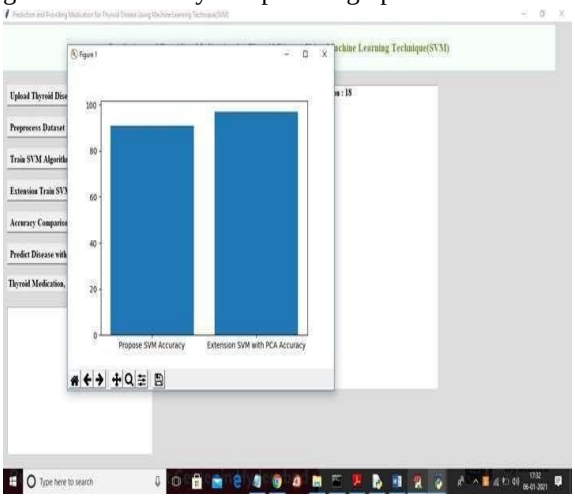

In above graph 84.15% and 6.81% is the true prediction and now clicks on 'Extension Train SVM with PCA Features Optimization' button to train SVM with PCA features optimization and to get below prediction accuracy



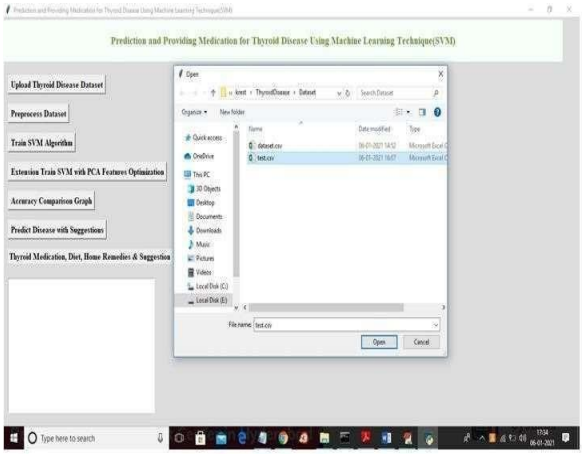

In above screen SVM with PCA extension got 96.98% prediction accuracy and confusion matrix values is also better compare to normal SVM and below is extension SVM confusion matrix graph



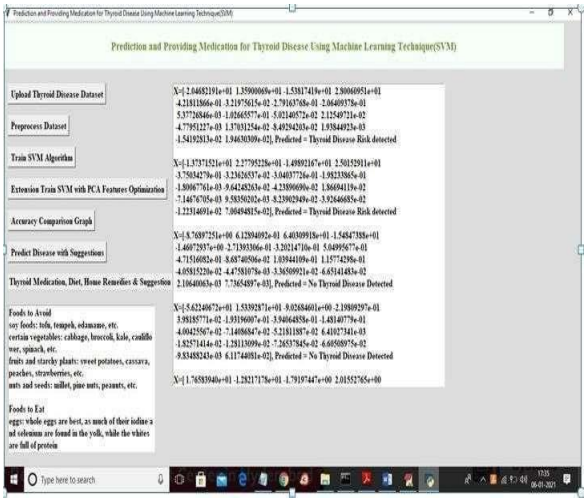

In above graph 89.70 and 7.29% is the correct prediction and other values are the false prediction. Now click on 'Accuracy Comparison Graph' button to get below accuracy comparison graph




In above graph x-axis represents algorithm name and y-axis represents accuracy of those algorithms and from above graph we can conclude that extension SVM with PCA is better than normal SVM and now click on 'Predict Disease with Suggestions' button to upload new test data and predict whether new test data contains thyroid or not.




In above screen selecting and uploading 'test.csv' file and then click on 'Open' button to upload test dataset and to predict disease and to get below screen




In above screen in brackets, we can see each record test value and after bracket we can see value as thyroid risk detected or not and if detected then it left box, we are showing diet and medication plan as suggestion.

## VI. CONCLUSION

The research work further studies the unusual machine learning strategies which can be mobilized in the diagnosis of thyroid diseases. In recent years, numerous approachable analyses for adequate and

professional diagnosis of thyroid disease have been developed and used. Study reveals that various technologies used in both articles demonstrating different precision. Most academic papers indicate that the neural network outperforms other strategies. In other hand, also due to the fact that the help vector machine and decision tree have done well. There is no question that experts around the world have gained a great deal of improvement in diagnosing thyroid diseases, but it is recommended that number of criteria used by patients to diagnose thyroid disorders can be limited. More characteristics mean that a patient needs to perform a larger range of health evaluations that are both cost-effective and time-consuming. Thus, certain algorithms and predictive models of thyroid disease need to be developed, requiring a minimum number of criteria for person to diagnose thyroid disease and saving patient's time and money.

## REFERENCE


[1] L. Ozyılmaz and T. Yıldırım, "Diagnosis of thyroid disease using artificial neural network methods," in: Proceedings of ICONIP'02 9th international conference on neural information processing (Singapore: Orchid Country Club, 2002) pp. 2033–2036.

[2] K. Polat, S. Sahan and S. Gunes, "A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted pre-processing for thyroid disease diagnosis," Expert Systems with Applications, vol. 32, 2007, pp. 1141-1147.

[3] F. Saiti, A. A. Naini, M. A. Shoorehdeli, and M. Teshnehlab, "Thyroid Disease Diagnosis Based on Genetic Algorithms Using PNN and SVM," in 3rd International Conference on Bioinformatics and Biomedical Engineering, 2009. ICBBE 2009.

[4] G. Zhang, L.V. Berardi, "An investigation of neural networks in thyroid function diagnosis," Health Care Management Science, 1998, pp. 29-37. Available: http://www.endocrineweb.com/ thyroid.html, (Accessed: 7 August 2007).

[5] V. Vapnik, Estimation of Dependences Based on Empirical Data, Springer, New York, 2012.

[6] Obermeyer Z, Emanuel EJ. Predicting the future— big data, machine learning, and clinical medicine. N Engl J Med. 2016; 375:12161219.

[7] Breiman L. Statistical Modeling: the two cultures. Stat Sci. 2001; 16:199-231.

[8] Ehrenstein V, Nielsen H, Pedersen AB, Johnsen SP, Pedersen L. Clinical epidemiology in the era of big data: new opportunities, familiar challenges. Clin Epidemiol. 2017; 9:245-250

[9] Ghahramani Z. Probabilistic machine learning and artificial intelligence. Nature. 2015; 521: 452-459.

[10] Azimi P, Mohammadi HR, Benzel EC, Shahzadi S, Azhari S, Montazeri A. Artificial neural networks in neurosurgery. J Neurol Neurosurg Psychiatry. 2015; 86:251-256.

[11] Deo RC. Machine learning in medicine. Circulation .2015; 132: 19201930.

[12] P.C. Austin, J.V. Tu, J.E. Ho, D. Levy, D.S. Lee, Using methods from the data- mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes, J. Clin. Epidemiol. 66 (4) (2013)398–407.

[13] A.K. Pandey, P. Pandey, K.L. Jaiswal, A heart disease prediction model using Decision Tree, IUP J Comput. Sci. 7 (3) (2013) 43.

[14] S. Ismaeel, A. Miri, D. Chourishi, in: Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis, IEEE Canada International Humanitarian Technology Conference, 2015, pp. 1–3.

[15] L. Verma, S. Srivastava, P.C. Negi, A hybrid data mining model to predict coronary artery disease cases using non- invasive clinical data, J. Med. Syst. 40 (7) (2016) 1–7.